УДК 577.214.626+316.452

NETINFERENCE: ПРОГРАММЫ ДЛЯ АНАЛИЗА СТРУКТУРЫ И ДИНАМИКИ СЕТЕЙ

© 2013 г. **И.И. Титов[1, 2], А.А. Блинов[2], К.А. Рудниченко[3], П.В. Крутов[2], А.Л. Казанцев[2], А.И. Куликов[2, 3]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: titov@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[3] Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия

Поступила в редакцию 15 августа 2013 г. Принята к публикации 5 сентября 2013 г.

В работе представлен пакет компьютерных программ для анализа структурно-функциональной организации и эволюции во времени биологических, социальных и других сетей. Пакет позволяет исследовать как глобальную архитектуру сетей, так и их локальные свойства, при этом выявлять ключевые регуляторы и структурно-функциональные модули, а также проследить развитие сетей во времени. Работа пакета иллюстрирована на примере нескольких генных сетей, сети соавторства научных публикаций в области биологии и медицины, а также сети терминов и ключевых слов из этой же области знаний.

Ключевые слова: генная сеть, сеть соавторства, сеть научных терминов, структура сети, динамика сети, синхронная булева модель, компьютерный анализ.

ВВЕДЕНИЕ

Удобным способом представления сложных систем являются сети. Большинство биологических, социальных, технологических и других сетей не являются регулярными или случайными, а обладают сходной сложной архитектурой связей. Для устройства этих сетей характерны сильная кластеризация и малый диаметр, свойство так называемого «малого мира». В результате такие сети демонстрируют интересные динамические свойства (Newman, 2003).

Для понимания организации и функционирования столь сложных сетей необходимо исследование их архитектуры с разных сторон и на разных масштабах рассмотрения: изучение глобальной и локальной структуры, выявление модулей и ключевых элементов, моделирование динамики и эволюции. В статье представлен пакет компьютерных программ, направленных на решение этой задачи, его работа продемонстрирована на примере некоторых генных, социальных и словарных сетей.

МАТЕРИАЛЫ И МЕТОДЫ

Пакет реализован в виде трех программ. Первая – программа для анализа глобальной и локальной архитектуры сетей. Программа реализована на языке C# и производит следующие вычисления: рассчитываются глобальные характеристики сети – распределение вершин по связям и аппроксимация этого распределения степенной зависимостью, диаметр сети и глобальный коэффициент кластеризации C1. В качестве локальных характеристик сети рассчитываются локальный коэффициент кластеризации C2 и коэффициент корреляции по степеням вершин. Неслучайно часто повторяющиеся модули сети (мотивы) находятся с помощью алгоритма

FANMOD (Wernicke, Rasche, 2006). Для ускорения расчетов на сетях больших размеров этот алгоритм модифицирован и использует библиотеку изоморфных графов, которая построена при помощи алгоритма Nauty (McKay, Piperno, 2013). Важно, что группы графов из библиотеки оказываются очень неравномерными по численности, что обосновывает необходимость точной оценки ожидаемой встречаемости при расчете статистической значимости мотива. В целом описанная программа носит общий характер и используется как дополнение для анализа конкретных сетей при помощи второй и третьей программ.

Вторая программа моделирует динамику и выявляет структурные модули и ключевые элементы генных сетей на основе синхронной булевой модели (Kauffman, 1969). Программа осуществляет полный перебор пространства состояний сети и определяет соседние во времени состояния, аттракторы, бассейны притяжения аттракторов и скорости переходов между бассейнами под действием шума заданной величины. Для ускорения расчетов и выявления структурно-функциональных организаций сети используются два подхода. В первом, статическом, сеть разбивается на полунезависимые кластеры одним из выбранных методов: «жадной» оптимизацией модульности, алгоритмом на основе определения смежности и случайных блужданий. После декомпозиции сети моделируется динамика каждого кластера по отдельности, затем на основе динамики отдельных кластеров восстанавливается динамический портрет всей сети. Во втором, динамическом, вершины сети начиная от полностью забуференных, рекурсивно удаляются в зависимости от степени их «канализированности» (Kauffman, 1969), пока не останется только «вычислительное ядро» сети, которое однозначно определяет динамику всей системы. В обратной процедуре динамика сети восстанавливается по динамике ее ядра. Влияние шума экспрессии генов на динамику сети моделируется методом Монте-Карло, в результате чего производится классификация вершин по степени их влияния на переходы между бассейнами притяжения стационарных состояний. Программа реализована на языке C++.

Третья программа реализована на языках SQL и Java и направлена на исследование ландшафта и временной эволюции тех сетей, которые могут быть получены из базы данных научных публикаций по биологии и медицине PubMed – сети соавторства и сети научных терминов. Кластеризация сети осуществляется при помощи модифицированного алгоритма SCAN (Xu *et al*., 2007). Восстановление эволюции сети осуществляется на основе определения соответствия между кластерами на соседних временных срезах. Периоды повышенного интереса к научной области определялись с помощью скрытой марковской модели для временного профиля частоты использования научных терминов.

РЕЗУЛЬТАТЫ

Выявление вычислительной архитектуры генных сетей на основе синхронной булевой модели

Тестирование программы рекурсивной редукции генной сети проводилось на сети ответа на стресс *E. coli*, исходно содержащей 73 вершины (Stepanenko, Titov, 2010), что соответствует 2^{73} возможным состояниям сети при полном переборе. Двукратное применение декомпозиции и редукции графа сократило его размер сначала до 32, а затем и до вполне вычислимого графа из 10 вершин, по состояниям которого восстанавливается полный динамический портрет сети.

Влияние шума экспрессии генов на динамику генной сети моделировалось для хорошо изученной генной сети морфогенеза цветка *Arabidopsis thaliana* (Alvarez-Buylla *et al*., 2008). Были выявлены критические динамические состояния генной сети, т. е. такие пары соседних состояний, которые принадлежат разным бассейнам. Информация о точках бифуркаций динамических траекторий генной сети была обобщена в виде ранжирования генов по степени влияния на неустойчивость траекторий (табл.). Из таблицы видно, что степень влияния шума может сильно варьировать от вершины к вершине. Более половины всех переходов между бассейнами были спровоцированы шумом в вершинах LFY и UFO. Первая из них соответствует гену Leafy, который инициирует развитие недифференцированных клеток, а вторая – F-box протеин, отвечающий за дифференциацию аттракторов Pet1–Pet2 и

Таблица

Относительная роль вершин, шум в которых инициирует переходы между бассейнами притяжения

Вершины генной сети, морфогенеза цветка	AG	AP1	AP2	AP3	EMF1	Ft	FUL	LFY	P1	Sep	TFl1	UFO	WUS
% от общего числа переходов	9,2	7,0	5,3	7,1	4,8	0,7	0	22,1	0,3	0,3	8,6	34,4	0

Stm1–Stm2. Известно, что мутагенное влияние на эти гены влечет различные нарушения в процессе развития цветка и приводит к изменению его внешнего вида.

Моделирование динамики генной сети методом Монте-Карло при заданном уровне шума позволяет построить матрицу Маркова. Эта матрица используется для определения кинетики системы в терминах населенности бассейнов (Alvarez-Buylla *et al*., 2008): порядка и скоростей переходов, устойчивости бассейнов и крупнозернистых кинетических мод, определяемых собственными числами матрицы.

Исследование развития научных коллективов ИЦиГ СО РАН на основе анализа сети соавторов научных публикаций

Построение временных срезов статистических характеристик сети и численности коллективов соавторов ИЦиГ СО РАН показывает, что, несмотря на нестационарный характер их поведения, можно выделить наиболее равномерный период развития, относящийся к 1998–2004 гг. (рис. 1, 2). При этом на протяжении всего рассмотренного времени наблюдаются возникновение, слияние, разделение и исчезновение кластеров соавторов (рис. 1). Наиболее существенное изменение структуры сети датируется 2007–2009 гг., сопряженными с выделением из института части лабораторий. Начиная с 2009 г. сеть возвращается к более плавной эволюции во времени (рис. 1, 2).

Построение динамики и ландшафта научных направлений на основе анализа сети научных терминов

В сравнении с сетями, которые были рассмотрены выше, еще более выраженно дисассортативными и кластеризованными (с высокими значениям коэффициента кластеризации и низким коэффициентом корреляции степеней вершин) оказались сети терминов научных публикаций. В качестве примера эволюции научной области

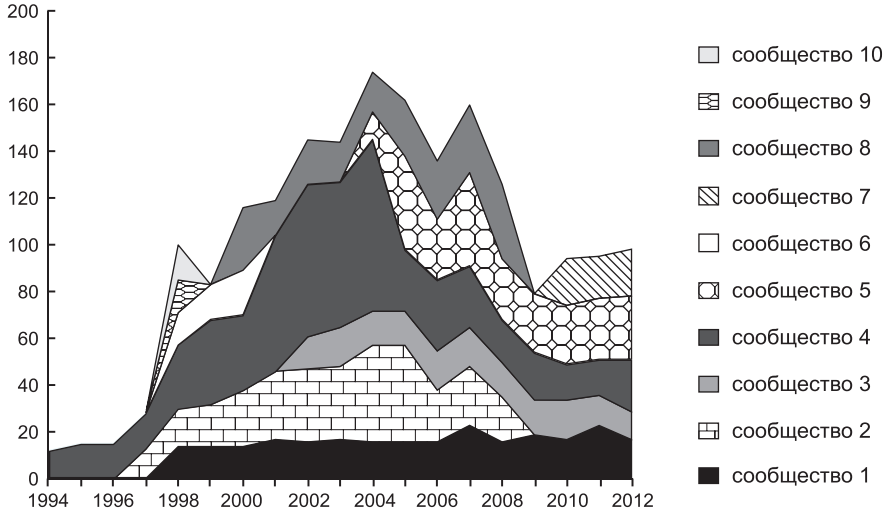

Рис. 1. Динамика численности наиболее крупных кластеров.

Каждый кластер соответствует сообществу внутри ИЦиГ СО РАН, которое образовано соавторством в научных публикациях, аннотированных в базе PubMed.

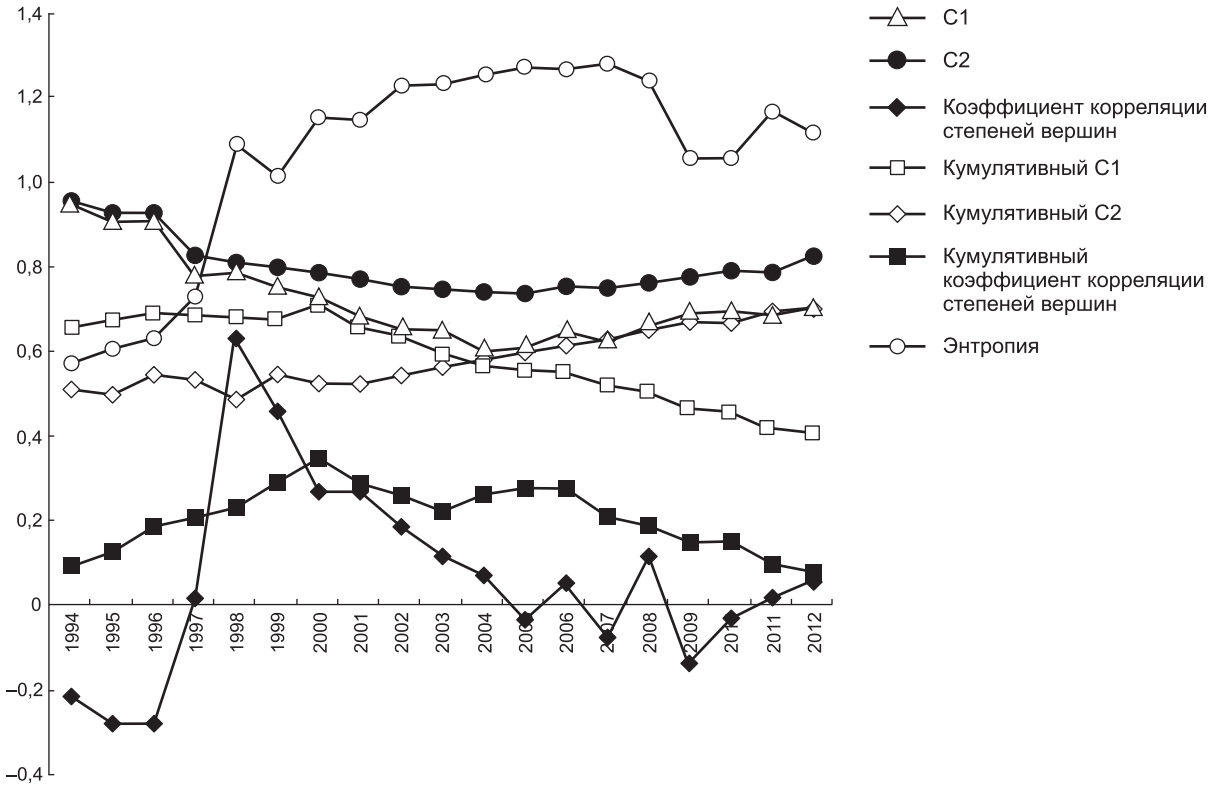

Рис. 2. Кумулятивные и мгновенные статистические характеристики сети соавторов ИЦиГ СО РАН.

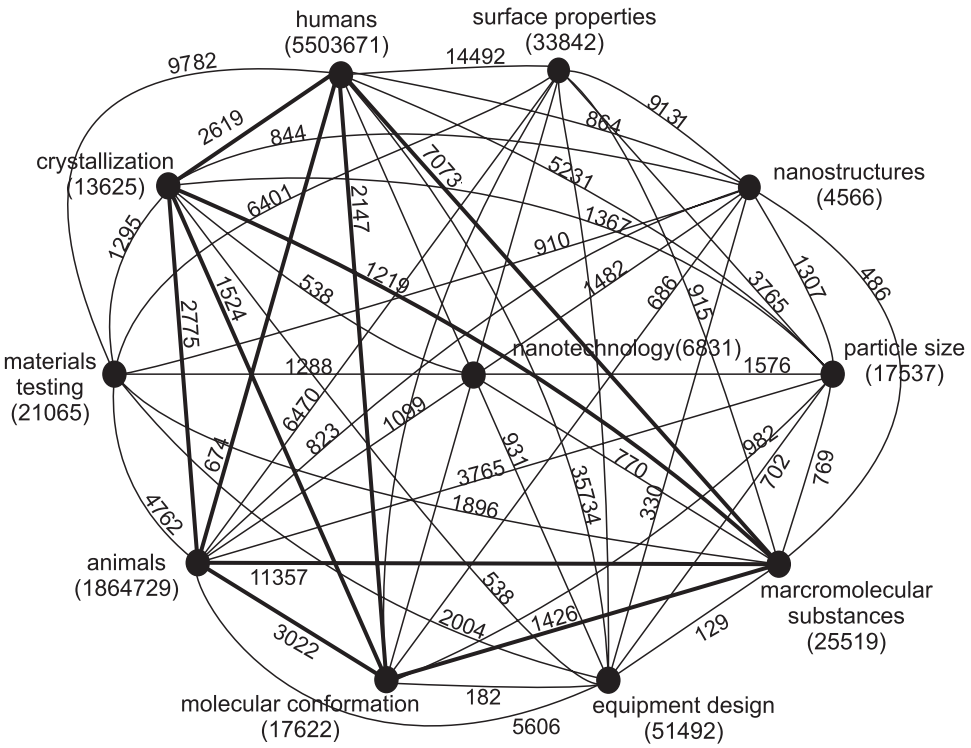

Рис. 3. Фрагмент кумулятивного ландшафта научных направлений в области нанотехнологий, построенного по аннотациям научных статей из базы PubMed.

Веса вершин соответствуют встречаемости терминов в аннотациях, веса ребер – совместной встречаемости терминов. Жирными ребрами показан кластер терминов, соответствующий исследованиям живой природы.

была рассмотрена область нанотехнологий. Хотя область науки «нанотехнология» имеет историю в несколько десятилетий и восходит к лекции Р. Фейнмана, само это слово возникло лишь около 30 лет назад. Еще более недавним является использование термина «nanotechnology» в кратком содержании научных статей в базе PubMed. Впервые термин употребляется в 1995 г., но с того момента частота его использования растет экспоненциально, что отражает взрывной всплеск интереса к этой области науки и ее присутствие в начальной стадии кривой Гартнера развития технологий. Построение сети терминов в области нанотехнологий показывает разделение на области знаний живой и неживой природы (рис. 3).

ЗАКЛЮЧЕНИЕ

Исследование сложных систем часто невозможно представить без изучения свойств сетей, моделирующих эти системы. Такие сети обычно обладают необычной топологией и характеризуются богатой динамикой. В работе представлен набор компьютерных программ для изучения биологических, социальных и других сетей. Разработанные программы предназначены для изучения глобальных и локальных свойств сложных систем, а также их развития во времени.

БЛАГОДАРНОСТИ

Работа поддержана Междисциплинарным интеграционным проектом СО РАН № 21 и Президентской программой по государственной поддержке ведущих научных школ РФ НШ-5278.2012.4.

ЛИТЕРАТУРА

Alvarez-Buylla E.R., Chaos A., Aldana M. *et al*. Padilla-longoria floral morphogenesis: stochastic explorations of a gene network epigenetic landscape // PLoS ONE. 2008. V. 3. No. 11.

Kauffman S.A. Metabolic stability and epigenesis in randomly constructed genetic nets // J. Theor. Biol. 1969. V. 22. P. 437–467.

McKay B.D., Piperno A. Practical graph isomorphism. II. 2013. 22 p. http://arxiv.org/abs/1301.1493.

Newman M.E.J. The structure and functions of complex networks // SIAM Rev. 2003. V. 45. No. 2. P. 167–256.

Stepanenko I.L., Titov I.I. Computer analysis of stress response network *E. coli* // Proc. 7th Int. Conf. on Bioinformatics of Genome Regulation and Structure\Systems Biology BGRS\SB.10. Novosibirsk, Russia, June 20–27 2010. Novosibirsk. P. 278.

Wernicke S., Rasche F. FANMOD: a tool for fast network motif detection // Bioinformatics. 2006. V. 22. No. 9. P. 1152–1153.

Xu X., Yuruk N., Feng Zh., Schweiger T.A.J. SCAN: a structural clustering algorithm for networks // Proc. KDD '07 Proc. of the 13th ACM SIGKDD Intern. Conf. on Knowledge discovery and data mining. N.Y., 2007. P. 824–833.

NETINFERENCE: COMPUTER PROGRAMS FOR REVEALING NETWORK STRUCTURE AND DYNAMICS

I.I. Titov[1,2], A.A. Blinov[2], K.A. Rudnichenko[3], P.V. Krutov[2], A.L. Kazantsev[2], A.I. Kulikov[2,3]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: titov@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia;
[3] Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia

Summary

We present a computer package for analyzing the structure-functional organization and evolution of biological, social and other networks. The programs allows investigation of not only the global network architecture, but also its local properties, revealing key regulators and structure-functional modules. Also, the network evolution can be traced. The package has been tested with two gene networks: the co-authorship network of biomedical papers and the biomedical term network.

Key words: gene network, co-authorship network, term network, network structure, network dynamics, synchronous Boolean model, computer analysis.